Desenvolvimento de marcadores moleculares espécie-específicos para a identificação de *Eucalyptus*

Hernando Javier Rivera Jiménez

Botucatu-SP
2016

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio De Mesquita Filho"

INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

Desenvolvimento de marcadores moleculares espécie-específicos para a identificação de *Eucalyptus*

Hernando Javier Rivera Jiménez

Orientador: Prof. Dr. Celso Luís Marino

Tese apresentada ao Instituto de Biociências, *Campus* de Botucatu, UNESP, para obtenção do título de Doutor no Programa de Pós-graduação em Ciências Biológicas: Genética.

Botucatu-SP
2016

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.

DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE - CRB 8/5651

Rivera-Jiménez, Hernando Javier

Desenvolvimento de marcadores moleculares espécie-específicos para a identificação de *Eucalyptus* / Hernando Javier Rivera Jiménez. – Botucatu, 2016

Tese (doutorado) - Universidade Estadual Paulista, Instituto de Biociências de Botucatu

Orientador: Celso Luis Marino

Assunto CAPES: 20203004

Aos meus pais Nohora Jimenez Sossa e (Hernando W. Rivera Acuña -in memoriam).
Obrigado Deus por colocá-los tão caprichosamente em minha vida.

Dedico.

AGRADECIMENTOS

Desafio tão grande quanto escrever esta tese, foi utilizar apenas duas páginas para agradecer as pessoas que fizeram possível minha formação doutoral.

Ao Programa de bolsa para Apoio a Estudantes do Exterior PAEDEX/ AUIP da Universidade Estadual Paulista (UNESP) de Brasil, pela bolsa de doutorado concedida.

Ao Programa de COLCIENCIAS - Colômbia por seu apoio no programa de doutorando no exterior.

Muito especialmente, desejo agradecer ao meu orientador Prof. Dr. Celso Luis Marino, pela disponibilidade, atenção dispensada, paciência, dedicação e profissionalismo ... um Muito Obrigado.

Ao Dr. Bruno César Rossini, pela colaboração e auxílio nas análises dos experimentos.

À empresa Suzano Papel e Celulose, especialmente ao Shinitiro, Esteban e Izabel, por fornecer o material vegetal estudado e colaboração no trabalho.

Instituto de Pesquisas e Estudos Florestais – IPEF, especialmente ao Dr. Paulo Muller Da Silva, por fornecer o material vegetal estudado e colaboração no trabalho.

Aos professores do Programa de Pós-Graduação em Ciências Biológicas (Genética), especialmente, à Prof. Dr. Claudia Aparecida Rainho, Prof. Dr. Adriane Pinto Wasko, Prof. Dr. Ligia Sousa e o Prof. Dr. Danillo Pinhal. Obrigado pelo exemplo e pela contribuição em minha formação acadêmica.

Aos funcionários da Seção Técnica da Pós-Graduação.

À minha Esposa Carolina por o tempo todo ao meu lado, incondicionalmente, nos momentos mais difíceis, meus filhos Julia Vanesa e Elías David por estar sempre ao meu lado me confortando.

À minha família, em particular a minha mãe Nohora Jimenez Sossa, meu irmão Mario A. Rivera J e sobrinho Marcos, por seu apoio incondicional, minha tia Nancy, meu primo Alexander, minha sogra Carmenza e cunhados por seu apoio.

Aos amigos do curso de pós-graduação, em especial à Leonardo Curi Martin, Lídia Carolina Arneiro, Karine Kettener, Maria Cecília Fuchs, Júlio C. Otto, Vanusa do S. Leite, Pedro Gabriel Nachtigall, Izabel Cristina, Swapnil Sanmukh pelo convívio, alegria, ajuda e paciência dedicados durante todo esse período.

Aos amigos Evandro V. Tambarussi, Saura Rodrigues, Michelle Pires, Danis Vivan, Luis Eduardo Paulista (Fagian), Wismar Sarmiento, Isamery Machado, Fernando Castro, Patrícia e Arturo Gomez Insuasti pelo apoio, lealdade, descontração e amizade dedicados desde sempre. Desculpem-me por não me arriscar a citar mas nomes, mas felizmente somos muitos e estamos sempre aumentando.

Enfim, a todos os amigos e familiares que, direta ou indiretamente, contribuíram para minha formação não só como cientista, mas principalmente como cidadão consciente, meus sinceros agradecimentos.

Para Julia Vanesa y Elías David, cuando nacieron florecieron los campos.

SUMMARY

List of Tables

Chapter I

Chapter II

List of Figures

Chapter I

Chapter II

Supporting Information Additional

Chapter II

PREFACE

The forest-breeding program in Brazil has the general objective of providing most adapted plants to different environments for various Brazilian regions, for fulfilling timber demands meant for multiple uses in the country. One of the main problems found in different forest breeding programs are the difficulty to identify the different species and hybrids. The use of molecular biology techniques in plant breeding programs is found very effective in the optimization of the time and the direction of these programs, particularly among those plants of the same subgenus. The process of selection and hybrid plants selected for planting in most cases; significantly increase the gain in terms of production and adaptability. The use of molecular markers to characterize the molecular variability of forest species has revolutionized genetic analysis in recent years. The bulk segregant analysis (BSA) is a technique used to identify molecular markers linked to monogenic, dominant or recessive characters. BSA technique in combination with Amplified Fragment Length Polymorphisms (AFLP) technique is an efficient methodology for the detection of polymorphism from genomic restriction fragments through PCR amplification; which helps in analyzing large number of loci for testing without the need for previous information of their sequence in respect to their dominance and reproducibility. The most recent and promising applications of molecular biological methods for the detection of small DNA fragments as identification tool and constituency of species in plants and animals is called DNA barcode. The use of barcode DNA sequence is useful for grouping data and analyzing jointly in order with ease of amplification and sequencing and quality of sequence discriminatory power of the marker.

The results obtained during the development of this study are presented in a scientific article format. The first paper was submitted to the Silvae Genetics journal, entitled **"Development of molecular markers for the eucalyptus species identification"**, which showed the analysis of AFLP markers and BSA that were associated with the identification of five species of *Eucalyptus* (*E. saligna*, *E. tereticornis*, *E. urophylla*, *E. grandis and E. brassiana*). The second article will be submitted to the Plos One journal with the title **"Evaluating the capacity of plant DNA barcodes to discriminate species of *Eucalypts*"**, we are presenting our results for evaluating of the discriminatory power of the DNA barcode label based on the use of internal transcribed spacer *ITS1, ITS2* and plastid genomes *rpoC1*, *matK, rbcL* and *ycf1*, enabling the genetic separation of 14 species of Eucalyptus.

Objectives

The general objective of this study was to evaluate the applicability of molecular markers associated with the identification of species of eucalypts.

They are also specific objectives:

Identifying by applying Bulk Segregant Analysis (BSA) and Amplified Fragment Length Polymorphism (AFLP) markers associated with specific species of *Eucalyptus*.

Evaluating and quantifying the DNA-barcode efficiency to discriminate species of *Eucalyptus*, based on internal transcribed spacer *ITS1, ITS2* and plastid genomes *rpoC1, matK, rbcL* and *ycf1*.

Chapter I

- Article submitted for journal *Silvae Genetica* (under revision).

Title: Development of molecular markers for the eucalypts species identification

Authored by H. J. Rivera-Jiménez*[1], B. C. Rossin[2], V. S. Leite[1], P. H. Muller Da Silva[3] and C. L. Marino[1].

Address:

1. Departamento de Genética, Instituto de Biociências, Universidade Estadual Paulista, Distrito de Rubião Junior, 18618-000, Botucatu – SP, Brasil.

2. Instituto de Biotecnologia / IBB- UNESP/, Botucatu, SP, Brasil.

3. Instituto de Pesquisas e Estudos Florestais, Av. Pádua Dias 11, C. P. 530, CEP 13400-970, Piracicaba, SP, Brasil. E-mail: paulohenrique@ipef.br .

*) Corresponding author: Departamento de Genética, Instituto de Biociências – UNESP.

Caixa postal 510 – CEP: 18618-970 – Botucatu – SP. Brazil, Phone: +55 (14) 3880-0780 - 38800391 E-mail: hriveraj@gmail.com; hriveraj@ibb.unesp.br

Keywords: *Eucalyptus*, AFLP, BSA, molecular marker

Word count for all text: 3,872

Abstract

One of the main problems faced in several *Eucalyptus* breeding programs is the difficulty to identify the species and hybrids. This study aimed to find molecular markers associated with five species of *Eucalyptus* (*E. saligna, E. tereticornis, E. urophylla, E. grandis* and *E. brassiana*), by AFLP (Amplified Fragment Length Polymorphism) markers and BSA (Bulk Segregant Analysis), for their use in breeding programs. In 33 primer combinations, a total of 868 polymorphic fragments was obtained, which represent a 91.65% of polymorphism. The best combinations that show potential markers for species identification were the primers M + GGT / E + ACC, which was linked to 70% of *E. urophylla* individuals. However, primer combination composed of M+GGA/E+ACC identified 60% of individuals in the *E. saligna* species; combination by the primers M+GTC/E+AAC, confirmed two marks, one in 60% and the other in 50% of *E. grandis* individuals in the identification test. The treatment composed by the primers M+GGC/E+AAA, was confirmed in only 30% of *E. brassiana* individuals, being the same for the combination M+GGC/E+ACC primers, identifying 30% of *E. tereticornis* individuals. The AFLP analysis and BSA provide a quick tool for the identification of cultivars in eucalypts and can also be used to assist forest breeding programs.

Introduction

Eucalyptus are planted on a large scale due to their high productivity obtained by breeding program that explored the adaptability of genus species to different environmental conditions (GONÇALVES et al., 2013). To know the species that are being worked in the breeding program is important for the strategic planning, especially to identify interspecific hybrids and germplasm conservation (MARCUCCI et al. 2003; BALLESTA et al, 2015). Overall, a good breeding program selects the best individuals in intrafamily level, increasing the plantations productivity performance and industrial products quality (GRATTAPAGLIA and KIRST, 2008).

Brazilian forest breeding programs aim to provide for all brazilian regions enough genetic variability in order to obtain plants adapted to different environments to attend the demand for wood (LEITE et al., 2011; FORRESTER and SMITH, 2012). However, one of the problems is the difficulty to identify the species, especially in hybrid combinations (GRATTAPAGLIA

et al., 2004). In this context, the incorporation of molecular biology techniques in plant breeding programs is allowing the optimization of time and the conduction of these programs, being the hybrid plants, selected for planting most of the time (PONGITORY et al., 2004; SOARES et al., 2010). In *Eucalyptus* the hybridization strategy can significantly increase production and adaptability in the resulting progeny, while the hybrid plants are commonly selected to be part of the commercial plantation, selection based on phenotypic superiority and genetic stability (POTTS and DUNGEY, 2004).

A tool to assist the breeding programs are the molecular markers. Bulked segregant analysis (BSA) is a method introduced by MICHELMORE et al. (1991), which has been used to identify molecular markers linked to a monogenic, dominant or recessive trait. The technique consists in comparing two sets of DNA samples from a segregating population, where each bulk is composed of individuals from the same species that have the same trait or gene of interest (BLANCO and VALVERDE, 2005; FUCHS et al., 2011). The AFLP (Amplified Fragment Length Polymorphism) markers are high efficient for detecting polymorphism of genomic restriction fragments by PCR amplification (polymerase chain reaction) (VOS et al., 1995). These markers are applied in genetic diversity studies in germplasm banks (RIVERA-JIMENEZ et al., 2011) and allow us to obtain a large number of tags randomly distributed in the genome (VOS et al., 1995).

This study aimed to identify molecular markers associated with five species of *Eucalyptus* through AFLP molecular markers and bulked segregant analysis (BSA).

Material and Methods

A total of 10 individuals from five *Eucalyptus* species was provided by Suzano Papel e Celulose SA (Table 1).

Table 1. - Proceedings of *Eucalyptus* species studied.

Species	Origin	Number of plants
Eucalyptus brassiana	Embrapa – CSIRO 10972 (North Moreton, QLD, Austrália)	10
Eucalyptus saligna	Coffs Harbour (Austrália)	10
Eucalyptus grandis	Coffs Harbour (Austrália)	10
Eucalyptus urophylla	IPEF – Timor	10
Eucalyptus tereticornis	Embrapa – CSIRO 10975-8140 (Cooktown e Laura, QLD, Austrália)	10

Genomic DNA was extracted according to the CTAB protocol described by DOYLE and DOYLE (1990), with some modifications as follow: 5% CTAB, removing the proteinase K extraction buffer; CIA step (chloroform: isoamyl alcohol 24: 1), was carried out only once; and finally, it was removed the cleaning step with NaCl to extract the DNA from fresh leaves. For each sample, approximately 50 mg of fresh leaf tissues was macerated without main vein. Quantification was performed in spectrophotometer Nano Drop®- ND1000. The DNA used in the amplification reactions was absent of impurities and phenolic compounds, diluted to a concentration of 50 ng/µl in autoclaved ultra pure water.

We worked with *E. saligna, E. tereticornis, E. urophylla, E. grandis* and *E. brassiana* species, building a bulk for each specie composed of 10 individuals, in a DNA concentration of 10 ng/µL per individual and the final concentration of each bulk were 100 ng/µL. The DNA bulks were screened for polymorphic markers using AFLP markers in order to identify polymorphisms associated with each specie.

AFLP protocol was adapted from VOS et al. (1995). Genomic DNA was digested with a combination of two enzymes, EcoRI + MseI. 700 ng of DNA were digested with 5U per enzyme, 5 uL of One Phor All buffer (OPA, Amersham), 0.5 uL of BSA (10 ug / uL) in a final volume 50 uL. The reaction was incubated at 37 ° C for 16 hours. Amplified products were visualized after electrophoresis in 0.8 % agarose gel in 1× TBE (Tris-borate-EDTA) stained with ethidium bromide and visualized under UV light transilluminator. The EcoRI

adapter was diluted to 5 pM solution containing 0.5 x One Phor All buffer 10x (OPA). The MseI adapter was diluted to 50 pM solution 0.5 x One Phor All buffer 10 x (OPA). The hybridization of adapters was performed in a thermocycler model PTC-100 (MJ Research ®) in a reaction consisted of 10 min 65 ° C, 10 min at 37 ° C and 10 min at 25 ° C. The adapters were ligated to the DNA fragments in a reaction containing 1 uL of the enzyme T4 DNA ligase buffer (10x), 1 uL of each adapter (5 or 50 pM), 3U T4 DNA ligase (Invitrogen, Carlsbad CA, USA), 6.67 uL of ultrapure water and 45 uL digested DNA solution. Ligation was performed at 17 ° C for 17 hours. There were four combinations of primers in the pre-amplification reactions (Ea / Mg and Ea / Mc), primers with complementary sequences to each of the adapters plus one selective nucleotide at the 3 'end (Table 2). The reactions were composed with 1 uL of each primer (25 ng / uL), 10 uL PCR Master Mix (Promega ®), 6 uL Nuclease-Free Water and 2 uL digested and ligated DNA in a final amount of 20 uL. The program for preamplification was: 94 ° C for 2 min, 26 cycles of 94 ° C for 1 minute, 56 ° C for 1 minute, 72 ° C for 1 minute and a final extension at 72 ° C for 5 minutes. In the product of this reaction was added 80 uL of ultrapure water.

Table 2. - Sequence of the adapters and primers used in the binding reactions, pre-amplification and amplification of the AFLP technique.

Adaptor or primer	**Oligonucleotide**
EcoRI adaptors	5′ CTCGTAGACTGCTACC 3′
	5′ AATTGGTACGCAGTCTAC 3′
Pre-selective amplification primer N: T ou A	5′ GACTGCGTACCAATTCN 3′
Selective amplification primer NNN: AAA,	5′ GACTGCGTACCAATTCNNN 3′
MseI adaptors	5′ GACGATGAGTCCTGAG 3′
	5′ TACTCAGGAACTCAT 3′
Pre-selective amplification primer N: T ou A	5′ GATGAGTCCTGAGTAAN 3′
Selective amplification primer NNN: AAA,	5′ GATGAGTCCTGAGTAANNN 3′

We tested 33 primers combinations for selective amplification (Table 3). In these reactions were used primers with sequences containing more three selective nucleotides at the 3 'end (Table 2), composed of 1 uL of each primer (25 ng / uL), 10 uL PCR Master Mix (Promega ®), 6 uL Nuclease-Free Water and 2 uL of pre-diluted reaction, in a final volume of 20 uL. The conditions for selective amplification was: 94 ° C for 2 min, 12 cycles of 94 ° C for 30 seconds, 65 ° C for 30 seconds and 72 ° C for 1 min, 23 cycles of 94 ° C for 30 seconds, 56 ° C for 30 seconds and 72 ° C for 1 min, with final extension at 72 ° C for 2 minutes. After that,

were added 8 uL of loading buffer (10 uL of formamide, 200 of 0.5M EDTA pH 8.0, 10 mg of bromophenol blue and xylene cyanol 10 mg). The samples were denatured for 5 minutes at 94 ° C and visualized after electrophoresis in a 6% polyacrylamide gel, 0.5 mm thick with the System "Sequi-Gen GT" (BioRad ®) 38 x 50 cm, stained with silver nitrate according to the protocol proposed by Hast et al. (2001). In cases of polymorphic fragments, their size was estimated by comparison with standard molecular weight of 100 bp (Promega ®).

In the band analysis, we considered especially those located between 200 and 700 bp. The bands located in the associated group in each species were taken as the model for the species molecular characterization. For each combination of primers, the amount of polymorphic bands and the percentage of each combination was established in the five bulk. In each of the combinations tested was taken as polymorphism any band that was different compared to the species group.

Table 3. - Combinations of primers used to obtain *Eucalyptus* AFLP markers.

Treatment	primer combinations		Treatment	primer combinations		Treatment	primer combinations	
	MSeI	*EcoRI*		*MSeI*	*EcoRI*		*MSeI*	*EcoRI*
1	GTG	AGA	12	CAA	AAC	23	GAA	AAA
2	GTG	AGC	13	CAA	AAA	24	GAA	AAC
3	GTG	AGG	14	CAT	AAA	25	GAA	ACC
4	GTG	ACA	15	CAT	AAC	26	GGC	AAA
5	GTT	AGC	16	CAT	ACC	27	GGC	AAC
6	CAG	AAC	17	CCT	AAA	28	GGC	ACC
7	CAG	AAA	18	CCT	AAC	29	GGT	AAA
8	CAG	ACC	19	CCT	ACC	30	GGT	ACC
9	CAC	AAC	20	GTC	AAA	31	GGA	AAA
10	CAC	AAA	21	GTC	AAC	32	GGA	AAC
11	CAC	ACC	22	GTC	ACC	33	GGA	ACC

Results

Based on the DNA mixture of individuals of each species was generated an AFLP profile of the five bulks by 33 primer combinations, resulting in 803 polymorphic fragments with 91.65% of polymorphism, the minimum and maximum number of fragments per primer was

nine and 62 fragments, respectively. Figure 1 and 2 shows the pattern of fragments formed by the DNA bulk and polymorphism found by BSA method. Considering each AFLP fragment as an independent locus, 868 different fragments were analyzed. The standard AFLP bands proved to be consistent and highly reproducible.

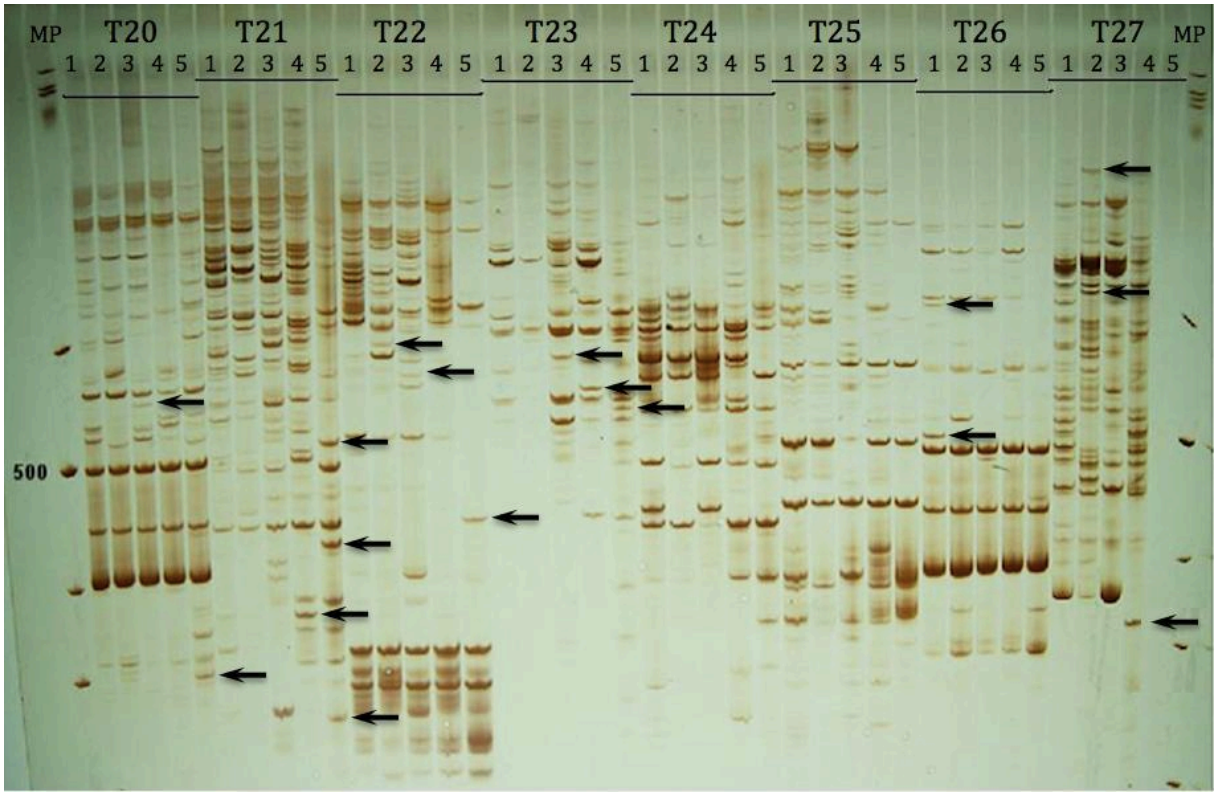


Figure 1. - Banding pattern formed by bulk DNA. They are suitable candidate markers associated with specific species in *Eucalyptus*. *Rail 1= *E. brassiana*, rail 2= *E. saligna*, rail 3= *E. tereticornis*, rail 4= *E. europhylla*, rail 5=*E. grandis.* MP= DNA Ladder 1 Kb.

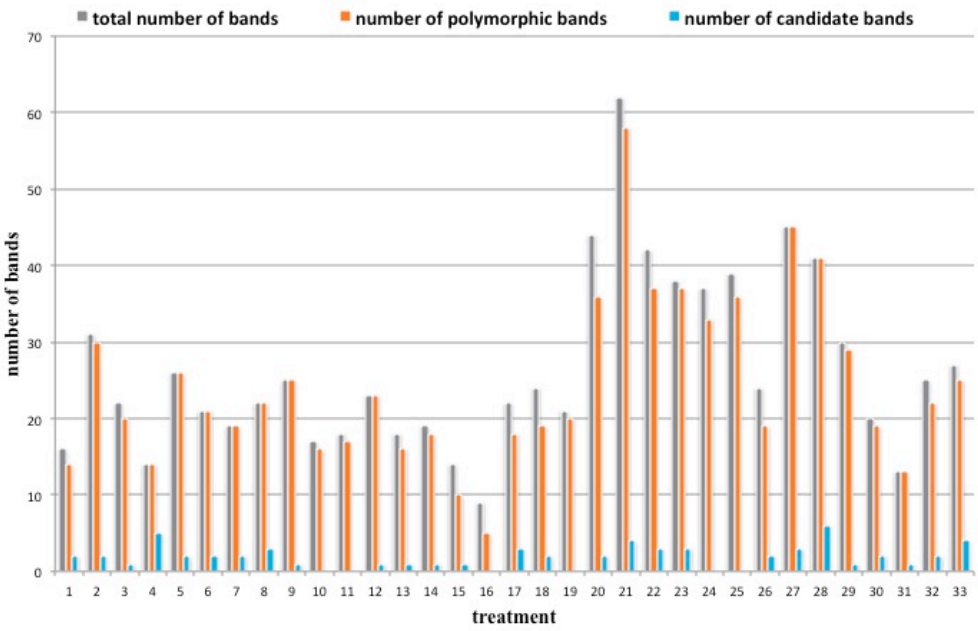


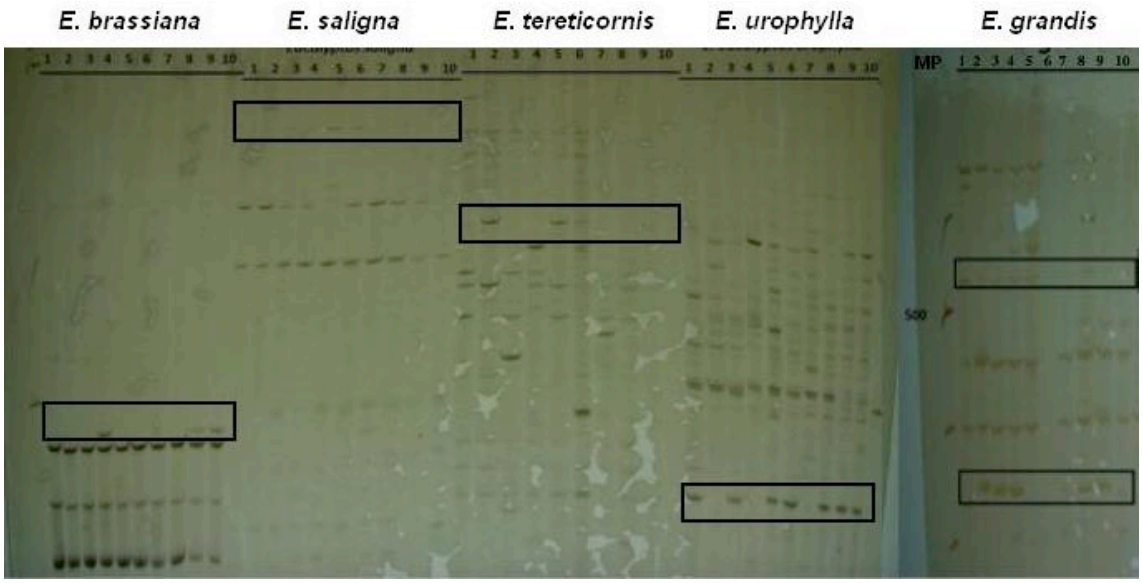
Figure 2. - Total primers and polymorphisms found in the BSA. Each combination was generated by *EcoRI* and *MseI* primer carrying a specific sequence of 3 nucleotides. The numbering corresponds to the number of specific primers shown in Table 1.

A total of 67 markers was identified to be *Eucalyptus* specie specific, raging between 200 bp and 700 bp, according to each primer set. It was found seven candidate fragments for *E. brassiana,* 18 fragments for *E. saligna*, 12 for *E. tereticornis*, 13 fragments for *E. urophylla* and 17 fragments for *E. grandis* (Table 4). In the gel for treatment 21 to 27, candidate marks were identified by bulk DNA to identify the species under study.

Table 4. - Primer combinations used for obtaining bands associated with the identification of *Eucalyptus* species.

Treatment	primer combinations		DNA Bulk					candidate fragments
	MseI primer	*EcoRI primer*	sp1/ *E. brassiana*	sp2/ *E. saligna*	sp3/*E. tereticornis*	sp4/*E. europhylla*	sp5/*E. grandis*	
1	GTG	AGA	0	1	1	0	0	2
2	GTG	AGC	0	2	0	0	0	2
3	GTG	AGG	0	0	0	0	1	1
4	GTG	ACA	0	2	3	0	0	5
5	GTT	AGC	0	2	0	0	0	2
6	CAG	AAC	0	1	0	0	5	6
7	CAG	AAA	0	0	0	2	0	2
8	CAG	ACC	0	0	1	1	1	3
9	CAC	AAC	0	0	0	0	1	1
10	CAC	AAA	0	0	0	0	0	0
11	CAC	ACC	0	0	0	0	0	0
12	CAA	AAC	0	0	0	0	1	1
13	CAA	AAA	0	0	1	0	0	1
14	CAT	AAA	0	0	0	1	0	1
15	CAT	AAC	0	1	0	0	0	1
16	CAT	ACC	0	0	0	0	0	0
17	CCT	AAA	0	1	1	0	1	3
18	CCT	AAC	0	0	0	0	2	2
19	CCT	ACC	0	0	0	0	0	0
20	GTC	AAA	0	0	1	0	1	2
21	GTC	AAC	0	0	0	1	3	4
22	GTC	ACC	0	1	1	1	0	3
23	GAA	AAA	0	0	1	1	1	3
24	GAA	AAC	0	0	0	0	0	0
25	GAA	ACC	0	0	0	0	0	0
26	GGC	AAA	2	0	0	0	0	2
27	GGC	AAC	0	2	0	1	0	3
28	GGC	ACC	2	0	0	4	0	6
29	GGT	AAA	0	0	1	0	0	1
30	GGT	ACC	1	0	0	1	0	2
31	GGA	AAA	0	1	0	0	0	1
32	GGA	AAC	1	2	0	0	0	3
33	GGA	ACC	1	2	1	0	0	4
Total			**7**	**18**	**12**	**13**	**17**	**67**

Of the 33 primers tested, the best combinations for species identification was T30 treatment, composed of M + GGT / E + ACC primer, which detected 70% of *E. urophylla* individuals. The treatment T33 (M + GGA / E + ACC primer) identify 60% of *E. saligna* individuals; the T21 treatment (M+GTC/E+AAC primer), had two markers, one 60% and the other 50% in the identification of *E. grandis* individuals. In T26 treatment (M+GGC/E+AAA primer), there was only confirmed 30% of *E. brassiana* individuals, and also for T28 treatment (M + GGC / E + ACC primer), was detected 30% of *E. tereticornis* individuals (Figure 3).



30%	60%	30%	70%	60-70%
T26	T33	T28	T30	T21

Figure 3. - Standard bands formed by DNA bulk open. For confirmation marks associated with specific species in *Eucalyptus*. *sp1= *E. brassiana,* sp2= *E. saligna*, sp3= *E. tereticornis*, sp4= *E. europhylla*, sp5=E. *grandis*. MP= DNA Ladder 1 Kb.

The combinations presented in this study show a high degree of polymorphism identifying candidate fragments for the species identification, with an average of 95.65%. However, once the bulk was opened, DNA detection power has decreased, from 30 to 70% of fragment frequency.

Discussion

The use BSA efficiency has been witnessed in other studies on the identification of markers linked to the genes that control resistance to rust on *Eucalyptus* sp. identifying polymorphism genetically linked to the characteristic resistance groups (ZAMPROGNO et al., 2008). LEITE et al. (2011) considers these primer combinations highly informative, being common and polymorphic in the *Eucalyptus* genome which makes these sequences valuable for the design of the other primer to generate polymorphisms, both for linkage maps or physical maps.

According to DOMINGUES et al. (2006), BSA technique allowed the identification of molecular markers linked to early flowering in *E. grandis* by RAPD (Random amplified polymorphic DNA) and SCAR (Sequence Characterized Amplified Region) with 60% of efficiency, confirming the usefulness of this technique as a molecular tool. These studies show that for an effective exploitation of bulked segregant analysis method, the only requirement is the existence of a segregating population of a gene of interest (MICHELMORE et al. 1991).

MELLISH et al. (2002), exploring the genetic diversity of the forage populations from the genus *Agropyron* spp. opted to use bulk and AFLP markers in identifying intra-population variation. Likewise, HERRMANN et al. (2005), worked with bulk by AFLP analysis to determine the genetic diversity and relationships within and among red clover populations (*Trifolium pratense* L.). The bulk analysis in white clover plants (*Trifolium repens* L.) by AFLP markers proved to be a powerful tool for the fast screening of genetic variability to identify cultivars (KÖLLIKER et al. 2001). According to MICHELMORE et al. (1991), BSA technique is a method that allows the identification of markers in specific genomic regions linked to any specific gene or genomic region. ZHU et al. (1998) raised some questions about the effectiveness of DNA bulks in relation to the low sensitivity to detect polymorphism using AFLP markers. However, in forest recently FUCHS et al. (2015) detected a marker totally linked genes associated with abnormal seedlings of eucalyptus who died in a few months by BSA technique.

However, when analyzed the bulks in studies with onion, barley, potato, lettuce, cabbage, and linen, the frequency of fragments was reduced below 50%, concluding that by using this method was not found a practical approach to detect genetic polymorphisms (VAN TREUREN, 2001). GUTHRIDGE et al. (2001) showed that profiles generated from individual samples could be or not present in the bulk samples, showing a high relationship between the frequency of occurrence and the presence of fragments in bulk samples.

Conclusions

The BSA method allowed to select a series of polymorphic AFLP markers associated with different eucalyptus species, confirming 70% of *E. urophylla* and 60% of *E. saligna* and *E. grandis* individuals. These molecular markers can be used as auxiliary tool to the rapidly and efficiently identification of *Eucalyptus* species.

Acknowledgments

We are grateful to E. V. TAMBARUSSI for assistance and discussions of this manuscript. We thank staff at the CAGEN-IBB for aid with marker analysis.

Conflicts of Interest

The authors declare no conflict of interest.

References

BALLESTA, P., F. MORA, R. I. CONTRERAS-SOTO, E. RUIZ and S. PERRET (2015): Analysis of the genetic diversity of *Eucalyptus cladocalyx* (sugar gum) using ISSR markers. Acta Scientiarum. Agronomy **37**: 133-140.

BLANCO, M. and R. VALVERDE (2005): Analisis de segregantes agrupados (BSA) para la deteccion de AFLPS ligados al GEN de resistencia a PVX *Solanum commersonii*.

Agronomia Costarricense **29**: 45–55.

DOMINGUES, D. S., A. P. CAZERTA, V. E. COSTRATO, E. J. DE MELO, S. ODA and C. L. MARINO (2006): Identificação de marcador RAPD E SCAR relacionados ao caractere florescimento precoce em *Eucalyptus grandis*. Ciência Florestal **16**: 251–260.

DOYLE, J. J. and J. L. DOYLE (1990): Isolation of plant DNA from fresh tissue. Focus **12**: 13–15.

FORRESTER, D. I. and R. G. B. SMITH (2012): Faster growth of *Eucalyptus grandis* and *Eucalyptus pilularis* in mixed-species stands than monocultures. Forest Ecology and Management **286**: 81–86.

FUCHS, M.C.P., TAMBARUSSI, E.V., LOURENÇÃO, J.C. (2015). Annals of Forest Science (2015) 72: 1043. doi:10.1007/s13595-015-0502-9

FUCHS, M. C., J. C LOURENÇÃO, E. V TAMBARUSSI, T. M BORTOLOTO, S. ODA, F. TS. NOGUEIRA and C. L MARINO (2011): Genome characterization of a *Eucalyptus* natural mutant. BMC. doi:10.1186/1753-6561-5-S7-P65

GONÇALVES, J. L. D. M., ALVARES, C. A., HIGA, A. R., SILVA, L. D., ALFENAS, A. C., STAHL, J., EPRON, D. (2013). Integrating genetic and silvicultural strategies to minimize abiotic and biotic constraints in Brazilian eucalypt plantations. Forest Ecology and Management, 301, 6–27. doi:10.1016/j.foreco.2012.12.030

GRATTAPAGLIA, D. and M. KIRST (2008): Eucalyptus applied genomics: from gene sequences to breeding tools. The New phytologist **179**: 911–29.

GRATTAPAGLIA, D., V. J. RIBEIRO and G. D. S. P, REZENDE (2004): Retrospective selection of elite parent trees using paternity testing with microsatellite markers: an alternative short term breeding tactic for Eucalyptus. TAG Theoretical and Applied

Genetics **109**:192–199.

GUTHRIDGE, K. M., M. P. DUPAL, R. KÖLLIKER, E. S. JONES, K. F. SMITH and J. W. FORSTER (2001): AFLP Analysis of Genetic Diversity within and between Populations of Perennial Ryegrass (*Lolium Perenne* L.). *Euphytica* **122**: 191–201. doi:10.1023/A:1012658315290.

HERRMANN, D., B BOLLER, F. WIDMER and R. KÖLLIKER (2005): Optimization of Bulked AFLP Analysis and Its Application for Exploring Diversity of Natural and Cultivated Populations of Red Clover. *Genome / National Research Council Canada = Génome / Conseil National de Recherches Canada* **48**: 474–86. doi:10.1139/g05-011.

KÖLLIKER, R., E. S JONES, M. Z. Z JAHUFER and J. W FORSTER (2001): Bulked AFLP Analysis for the Assessment of Genetic Diversity in White Clover (*Trifolium Repens* L.). Euphytica: 305–315.

LEITE, V., J. S. OTTO, C. D. SAGAWA, E. GONZALEZ, M. FAGUNDES, S. ODA and C. L MARINO (2011): Identification of Genomic Regions Related to Early Flowering in Eucalyptus. *BMC Proceedings* **5** (Suppl 7): P53. doi:10.1186/1753-6561-5-S7-P53.

MARCUCCI P. S. N., N. ZELENER, J. RODRIGUEZ, P. GELID and H. E HOPP (2003): Selection of a seed orchard of *Eucalyptus dunnii* based on genetic diversity criteria calculated using molecular markers. Tree physiology **23**: 625–32.

MELLISH, A., B. COULMAN and Y. FERDINANDEZ (2002): Genetic Relationships among Selected Crested Wheatgrass Cultivars and Species Determined on the Basis of AFLP Markers. Crop Science **42**: 1662–1668.

MICHELMORE, R. W., I. PARAN and R. V. KESSELI (1991): Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to

detect markers in specific genomic regions by using segregating populations. Proceedings of the National Academy of Sciences of the United States of America **88**: 9828–9832, 1991.

PONGITORY, V., G. MOURA, E. BLUME, E. FLORES and D. I. JAYA (2004): O germoplasma de *Eucalyptus urophylla* S. T. BLAKE no Brasil. Comunicado técnico. Blake 1977. https://www.infoteca.cnptia.embrapa.br/handle/doc/174980. 19/03/2016.

POTTS, B. M. and H. S. DUNGEY (2004): Interspecific hybridization of Eucalyptus: key issues for breeders and geneticists. New Forests **27**: 115–138.

RIVERA-JIMÉNEZ, H. J., A. ALVAREZ, J. D. PALACIO-MEJIA, D. Y. BARRIOS and D. LOPEZ (2011): Diversidad genética intra e inter-específica de ñame (*Dioscorea* spp.) de la región Caribe de Colombia mediante marcadores AFLP. Acta Agronómica **60:** 328–338.

SOARES, N. S., M. L. DA SILVA, J. L. P. DE REZENDE and M. F. M. GOMES (2010): Competitividade da cadeia produtiva da madeira de eucalipto no Brasil. Revista Árvore **34**: 917–928.

VAN TREUREN, R. (2001): Efficiency of reduced primer selectivity and bulked DNA analysis for the rapid detection of AFLP. Euphytica **117**: 27–37.

VOS, P., R. HOGERS, M. BLEEKER, M. REIJANS, T. VANDELEE, M. HORNES, A. FRIJTERS, J. POT, J. PELEMAN, M. KUIPER and M. ZABEAU (1995): AFLP – a new technique for DNA fingerprinting. Nucleic Acids Res **23**: 4407-4414.

ZAMPROGNO, K. C., E. L. FURTADO, C. L. MARINO, C. A. BONINE and D. C. DIAS (2008): Utilização de Análise de Segregantes Agrupados Na Identificação de Marcadores Ligados a Genes Que Controlam a Resistência À Ferrugem (*Puccinia Psidii Winter*) Em

Eucalyptus Sp. *Summa Phytopathologica* **34**: 253–255. doi:10.1590/S0100-54052008000300009.

ZHU, J., M. D. GALE, S. QUARRIE, M. T. JACKSON and G. J. BRYAN (1998): AFLP Markers for the Study of Rice Biodiversity. *TAG Theoretical and Applied Genetics* **96**: 602–611. doi:10.1007/s001220050778.

Chapter II

Evaluating the capacity of plant DNA barcodes to discriminate species of *Eucalyptus*

Hernando Rivera-Jiménez*[1], Bruno C. Rossini[2] and Celso L. Marino[1, 2]

[1.] Department of Genetics, Institute of Biosciences, São Paulo State University. Distrito de Rubião Junior, 18618-000, Botucatu – SP, Brazil, [2.] Biotechnology Institute of UNESP / IBB-UNESP/, Botucatu, SP, Brazil.

*hriveraj@gmail.com; hriveraj@ibb.unesp.br

Abstract

The forest-breeding program in Brazil has the general objective of providing most adapted plants for different environments, the breeding programs have some trouble in species identification. DNA barcoding was expected to be an effective tool for species identification in *Eucalyptus*. Fourteen *Eucalyptus* species were selected from the forest-breeding program in Brazil and tested four regions in the plastid genome (*matK, rbcL, rpoC1* and *ycf1*), a nuclear transcribed spacer (ITS1+ITS2*,* herein ITS region) and their combinations, in order to discriminate them at species level. Among the evaluated loci, ITS*, rbcL*, *rpoC1, matK* had the highest success rate for amplification (100%), followed by a low percentage of success by *ycf1* (67.14%). The "best match" and "best close match" approaches revealed a rate of correct species identification *ycf1*+ITS+*matK* (62.16%) followed by *matK*+ITS1 (61.64%) loci. Neighbour-joining cluster analysis indicated the highest degree of the 25 possible

combinations of the five regions; three provided the highest degree of species resolution (78.6%). Among these, a combination of ITS+*rbcL,* ITS+*matK*+*rpoC1* and ITS+*rbcL*+*rpoC1*, which comprises two and three DNA regions, is the best option for barcoding of *Eucalyptus* species.

Introduction

In many industrial sectors there is a great demand for *Eucalyptus* wood because of its use in pulp and the paper and cellulose production, coal, resin, construction, latex, cosmetics and essences, it is considered the most planted forest genus around the world (1). The forest breeding program has the general objective of providing the various Brazilian regions with sufficient genetic variability in such a way that gets more plants adapted to different environments, to meet timber demands for multiple uses (2). However one of the main problems found in different forest breeding programs is the existing difficulty to identify the different species and hybrids. In this context, incorporation of molecular biology techniques in plant breeding programs is allowing the optimization of the time and the direction of these programs, especially among those of the same subgenus, considerably increasing the gain in terms of production and adaptability in the progeny resulting from the process of selection (1,3).

The latest and most promising applications of molecular biological methods for the detection of units is the use of small DNA fragments, DNA barcoding, as identification tool for species proposed by Hebert et al. (2003). In animals the mitochondrial gene Cytochrome oxidase subunit 1 (*COI* or *COX1*) is well established as DNA barcoding, butthe standardization plant still remains unclear (4–6). Significant progress has been make in the DNA barcoding of higher plants, and the following core DNA barcodes have been proposed: the nuclear internal

transcribed spacer (ITS1 and ITS2) and several plastid genome regions such as *atpF–atpH, trnH–psbA, psbK–psbI, matK, rbcL, rpoB, rpoC1* and *ycf1* that are frequently used in plant molecular identification, but the combination accepted by community as a standardized DNA barcode region for plants is the combination of *matK* and *rbcL* regions (7–10). Unfortunately, in several orders this approach is not suitable for differentiate the majority species, being necessary the use of others most variable regions (11). Phylogenetic studies of Myrtaceae genus was performed using the sequences of the nuclear ITS region and the sequences of chloroplast (*psbA-trnH*) (12). Moreover, Steane et al. (2007) suggested that non-coding regions (the nuclear ITS region intergenic spacer) have potential in *Eucalyptus* and other related genus (*Corymbia* and *Angophora*) in phylogenetic analysis, this region also to been used to support monophyletic groups in *Eucalyptus* species of the genus *Corymbia* (14).

Due to the great importance of developing new molecular markers for efficient and reliable way of *Eucalyptus* to help the different programs of forest breeding to identify species and hybrids allowing the optimization of time and direction of these programs, in this study, we assessed six candidate barcodes by sampling 14 species of *Eucalyptus* using various evaluation methods with the following aims: (1) propose a more practical and universal barcode for *Eucalyptus* and (2) test the effectiveness of DNA barcoding for the identification of *Eucalyptus* species.

Materials and Methods

Plant materials, DNA extraction, PCR amplification, sequencing and public sequences

We analyzed 14 *Eucalyptus* species that are maintained at Instituto de Pesquisas e Estudos Florestais (IPEF) and Suzano Papel e Celulose SA (Suzano). A minimum of five individuals of each of *Eucalyptus* species were sampled (Table 1). As a result, a total of 70 specimens

were analyzed. Young leaves (with petiole and stem) were detached from each plant, and a 1 cm^2 piece of leaf tissue was desiccated in an airtight plastic bag containing silica gel. A total DNA of was isolated using a modified CTAB 5% protocol (15). PCR products for two nuclear internal transcribed spacer (ITS1 and ITS2) and four plastid barcodes (the coding genes *rpoC1, matK, rbcL* and *ycf1*) were amplified and sequenced using universal primers (11,16–18) (Table 2). The selected DNA regions were amplified by PCR. The PCR mix (10 µL) contained approximately 50 ng (1 µL) of template DNA, 5µL of 2×PCR Mastermix (0.005 units/µL Taq DNA polymerase; 4mMMgCl2; and 0.4mM dNTPs, Promega), 0.6 µL (10 µM) of each primer and 2,8 µL of ddH2O. The sequencing reactions were performed in both directions using the Applied Biosystems Prism Bigdye Terminator Cycle Sequencing Kit V 3.0 (Applied Biosystems).

Table 1. Classification and origin (Brazil) of the genus *Eucalyptus* and *Corymbia* species analyzed in this study

Code	*Species*	*Genus*	Subgenus	*Section*	*Collection site*	*Origins*
He	*E. Corymbia henryi*	*Corymbia*		Ochraria	E. E. Itatinga-SP[1]	Lockyer (Austrália)
Ar	*E. argophloia*	*Eucalyptus*	*Symphyomyrtus*	*Adnataria*	E. E. Itatinga-SP[1]	SO bulk (Austrália)
Mol	*E. moluccana*	*Eucalyptus*	*Symphyomyrtus*	*Adnataria*	E. E. Itatinga-SP[1]	Ravenshoe (Austrália)
Ce	*E. crebra*	*Eucalyptus*	*Symphyomyrtus*	*Adnataria*	E. E. Itatinga-SP[1]	33.9K NW BARADINE PO (Austrália)
Bro	*E. brookeriana*	*Eucalyptus*	*Symphyomyrtus*	*Maidenaria*	E. E. Itatinga-SP[1]	OTWAYS (Austrália)
Mac	*E. macarthurii*	*Eucalyptus*	*Symphyomyrtus*	*Maidenaria*	E. E. Itatinga-SP[1]	Paddys River, N (Austrália)
Lo	*E. longirostrata*	*Eucalyptus*	*Symphyomyrtus*	*Latoangulatae*	E. E. Itatinga-SP[1]	STARKVALE CREEK (Austrália)
Ma	*E. major*	*Eucalyptus*	*Symphyomyrtus*	*Latoangulatae*	E. E. Itatinga-SP[1]	27K SE GYMPIE (Austrália)
S	*E. saligna*	*Eucalyptus*	*Symphyomyrtus*	*Latoangulatae*	Itatinga-SP[2]	Coffs Harbour (Austrália)
G	*E. grandis*	*Eucalyptus*	*Symphyomyrtus*	*Latoangulatae*	São Miguel Arcanjo-SP[3]	Coffs Harbour (Austrália)
U	*E. urophylla*	*Eucalyptus*	*Symphyomyrtus*	*Latoangulatae*	Angatuba- SP[4]	IPEF – Timor
T	*E. tereticornis*	*Eucalyptus*	*Symphyomyrtus*	*Exsertaria*	Açailandia- MA[5]	Embrapa – CSIRO 10975-8140 (Cooktown e Laura, QLD, Austrália)
B	*E. brassiana*	*Eucalyptus*	*Symphyomyrtus*	*Exsertaria*	Urbano Santos-MA[6]	Embrapa – CSIRO 10972 (North Moreton, QLD, Austrália)
Am	*E. amplifolia*	*Eucalyptus*	*Symphyomyrtus*	*Exsertaria*	E. E. Itatinga[1]	NERONG S.F. (Austrália)

1. E. E. Itatinga - Latitude: 23° 10' S Longitude: 48° 40' W Altitude: 857 m. IPEF; 2. Itatinga - Latitude: 22º 53´S Longitude: 48º26´W Altitude: 800m. Suzano company; 3. São Miguel Arcanjo - Latitude: 23º52´S Longitude: 47º59´W Altitude: 650 m. Suzano company; 4. Angatuba - Latitude: 23º47´S Longitude: 48º42´W Altitude: 650 m. Suzano company; 5. Açailandia - Latitude: 23º47´S Longitude: 48º42´W Altitude: 650 m. Suzano company; 6. Urbano Santos - Latitude: 3º12´S Longitude: 43º24´W Altitude: 50 m. Suzano company.

Primer design *matK* and *rbcL*

Initially we used the *matK* and *rbcL* primers described by Tokuoka & Tobe (2006), however these primers not successfully amplified, therefore new primers for this region were designed. A total of 34 *matK* and *rbcL* sequences the deposited in GenBank were downloaded and extracted from the plastid genomes *Eucalyptus* (Table S1). All these complete genomes of chloroplasts were manually adjusted with the Bioedit software (19). Alignments we made using the algorithm MUSCLE (Multiple Sequence Comparison by Log-Expectation) (20). These sequences were used as initial templates for the design of several primer pairs spanning the roughly identified regions for each group using Primer3 v 0.4.0 (21). The successfully amplified fragments were finally sequenced (Table 2).

Table 2. List of primers used for PCR and sequence in this study

Region	primer	Sequence 5′-3′	Tm (°C)	Reference
ITS1	5a fwd	CCTTATCATTTAGAGGAAGGAG	50	(17)
	4 ver	TCCTCCGCTTATTGATATGC		
ITS2	S2F	ATGCGATACTTGGTGTGAAT	56	(17)
	S3R	GACGCTTCTCCAGACTACAAT		
rpoC1	1 F	GTGGATACACTTCTTGATAATGG	53	(16)
	4R	CCATAAGCATATCTTGAGTTG		
*rbcL	fwd	CTTGGCAGCATTCCGAGTA	62	This study
	Rev	CGGCTTCGATCTTTTTCAAT		
*matK	fwd	GACAATGATCCAATCAGAGGAA	62	This study
	Rev	TCGAAAATGCAGGTTATGACA		
ycf1	ycf1bF	TCTCGACGAAAATCAGATTGTTGTGAAT	57	(11)
	ycf1bR	ATACATGTCAAAGTGATGGAAAA		

* Primer developed in this study

Data analysis

DNA barcodes candidates were edited with BioEdit software, version 7.0.9.0 (19). Informative polymorphic characters were identified by MEGA6 (22). Alignment of the sequences was executed by MUSCLE program (20). Eventually manual adjustments were made through BioEdit software. The different locus combinations were partitioned for independent model assessment at each marker. Sequence distances were computed using Neighbor-joining (NJ) analysis with Kimura 2-Parameter (K2P) model (23). The program TaxonDNA (24) was used to test the accuracy of species assignments, cluster analysis and distribution of interspecific and intraspecific distances in the dataset. This analysis determines the closest match of a sequence from comparisons with all other sequences in an aligned data set. The similarity threshold is then established on the frequency distribution of the intraspecific pairwise distances. The threshold is set at a value below which 95% of all intraspecific pairwise distances are found (24). To test the accuracy of species assignments of the samples, "best match", "best close match" and "all species barcodes" functions of the program were used. Number of clusters produced by each locus or a combination of loci at 1% and 0.5% threshold were determined by the 'cluster' function that identifies clusters of similar sequences. Although there is no 'cut-off' threshold reported in plant barcode literature for the species discrimination, our intent was to test the efficiency of each barcode or combination of barcodes to produce single-species clusters at a set threshold. Distribution of pairwise interspecific and intraspecific distances in the dataset for each locus or combination of loci was analyzed by the "pairwise summary" function of the program.

Consensus barcodes were used in NJ cluster analysis to visualize the patterns of sequence divergence among taxa using MEGA6 (22). Node support was assessed by bootstrap test (1000 replicates) (25). The trees were drawn to scale with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. We employed a tree-

based method to evaluate the degree of species resolution (identification). Each barcode region and possible combinations of the regions were evaluated for the degree of species resolution they provided. For each set of data, the degree of species resolution for barcode regions was calculated as a percentage of the total number of species correctly identified in the NJ tree divided by the total number of species examined in the study. Species were determined in the NJ tree by examining pairwise branch lengths; if two species were separated by a branch length greater than zero and bootstrap greater than 50%, they were considered as separate species. Although this method does not address the differences of insertion and deletion between sequences, it approximates the degree of species differentiation for a given DNA barcode region based on sequence similarity.

Results

PCR amplification and sequencing

The sequence information of five candidate DNA barcode markers, ITS*, rpoC1, matK, rbcL*, and *ycf1*, is provided in Table 3 (and in S1 File). Sequencing success rates were 87.14% (ITS), 88,57% (*rpoC1*), 87,14% (*matK*), 80% (*rbcL*) and 40% (*ycf1*). We used the complete ITS region (*ITS1-5.8S-ITS2*) as a single barcode locus. Unfortunately for *ycf1*, the universal primer proposed by Dong et al. (11) did not returned a great amplification success.

Table 3. Summary statistics for potential barcode loci from five *Eucalyptus* species

	ITS	rpoC1	*matK	*rbcL	ycf1
Universality of primers	Yes	Yes	Yes	Yes	Partial
Percentage PCR success (%)	100	100	100	100	67.14
Percentage sequencing success (%)	87.14	88.57	87.14	80	40
Length of aligned sequence (bp)	800	594	892	840	663
No. of parsimony informative sites/variable sites	94/138	8/29.0	32/85	25/77	71/190
No. of species samples (individuals)	81	64	81	67	26
Ability to discriminate (NJ)	57%	7.1%	28.5%	14.3%	43%

* Primer developed in this study

Intra-specific variation and inter-specific divergence

The aligned sequence lengths ranged from 892 bp for *matK* to 594 bp for *rpoC1* (Table 3 and S2 File). The ITS region had the most variable sites and parsimony-informative traits, followed by *ycf1* (Table 2). The pairwise intraspecific distances in the thirty barcodes (combined or not) ranged from a minimum of 0.0% to a maximum of 9.92% (Table 4). The mean intraspecific distances were the minimum for *rpoC1* (0.10%) and the maximum for *ycf1* (2.5%). The pairwise interspecific distances in the thirty barcodes ranged from a minimum of 0% to a maximum of 14.7% (Table 3). The mean interspecific distances were minimum for *rpoC1* (0.25%) and maximum for *ycf1* (4.5%), in summary, *ycf1* exhibited the highest mean intra- and interspecific distance. A combination of sequences of different barcode loci increased the intraspecific and interspecific mean distances (Table 4). The data showed obvious overlapping between intraspecific and interspecific distances of the individual or

combined sequences (Table 4). This overlapping was minimum (26.97%) in ITS sequences and maximum (94.34%) in the two-gene combination (*matk+rpoC1*) (Table 4).

Table 4. Summary of the pairwise intraspecific and interspecific distances in the barcode loci of *Eucalyptus* species.

Barcode locus	Intraspecific distances (%			Interspecific distances (%)			overlap with 5% error margin on both sides	
(no. of sequences)	Min	Max	Mean	Min	Max	Mean	Overlapping distance range	Intra-/interspecific sequences in the overlap
ITS (81)	0.0	2.58	0.73	0.0	11.6	3.00	0.80-1.83	26.97%
rpoC1 (64)	0.0	1.4	0.10	0.0	1.8	0.25	0.0-0.68	92.46%
matK (81)	0.0	4.15	0.33	0.0	8.08	0.84	0.0-1.57	91.75%
rbcL (67)	0.0	2.42	0.39	0.0	4.12	0.68	0.0-2.03	93.57%
ycf1(26)	0.0	7.00	2.5	0.0	14.7	4.5	0.64-7.07	71.07%
ITS+rbcL(72)	0.0	2.63	0.65	0.0	10.53	1.94	0.4-1.73	61.07%
ITS+rpoC1 (72)	0.0	1.48	0.17	0.0	10.5	1.98	0.46-1.78	61.99%
rbcL+rpoC1 (70)	0.0	1.49	0.30	0.0	3.42	0.50	0.0-1.05	89.74%
matK+ITS (73)	0.0	2.92	0.61	0.0	10.53	1.93	0.5-2.02	69.17%
matK+rpoC1 (72)	0.0	4.15	0.29	0.0	4.51	0.63	0.0-1.73	94.34%
matk+rbcL (70)	0.0	2.02	0.33	0.0	4.51	0.75	0.11-1.69	85.65%
ycf1+ITS (73)	0.0	3.09	0.96	0.0	11.71	2.97	0.88-2.51	60.38%
ycf1+matK (69)	0.0	4.15	0.57	0.0	9.36	1.07	0.0-3.91	93.22%
ycf1+rbcL (66)	0.0	2.96	0.62	0.0	9.36	0.92	0.0-2.32	89.83%
ycf1+rpoC1 (66)	0.0	6.22	0.53	0.0	12.93	0.68	0.0-2.73	93.21%
ITS+rbcL+rpoC1 (67)	0.0	2.13	0.33	0.0	10.13	1.53	0.35-1.27	54.54%
matK+rbcL+rpoC1 (72)	0.0	1.69	0.28	0.0	4.51	0.61	0.08-1.25	84.5%
ITS+matK+rpoC1 (73)	0.0	2.92	0.52	0.0	10.13	1.52	0.44-1.49	66.41%
matK+ITS+rbcL (73)	0.0	2.63	0.55	0.0	10.53	1.6	0.41-1.51	63.34%
ycf1+ITS+rbcL	0.0	2.55	0.72	0.0	10.53	2.04	0.47-2.04	66.15%
ycf+ITS+rpoC1	0.0	3.09	0.72	0.0	11.39	2.14	0.51-2.13	66.27%
ycf1+ITS+matK (74)	0.0	2.92	0.70	0.0	10.53	2.01	0.52-2.29	70.42%
ycf1+matK+rbcL (71)	0.0	2.55	0.47	0.0	9.36	0.89	0.12-1.89	82.67%
ycf1+matK+rpoC1 (73)	0.0	4.15	0.42	0.0	9.36	0.77	0.0-2.18	91.48%
ycf1+rpoC1+rbcL (71)	0.0	2.55	0.44	0.0	9.36	0.66	0.0-1.66	91.56%
matK+ITS+rbcL+rpoC1 (73)	0.0	2.63	0.50	0.0	10.13	1.33	0.36-1.15	57.78%
ycf1+ITS+matK+rbcL (69)	0.0	2.55	0.59	0.0	10.53	1.68	0.43-1.63	65.64%
ycf1+ITS+matK+rpoC1 (73)	0.0	9.92	0.59	0.0	10.13	1.60	0.45-2.04	72.51%
ycf1+matK+rbcL+rpoC1 (73)	0.0	2.55	0.40	0.0	9.36	0.73	0.08-1.47	83.05%
ycf1+ITS+matK+rbcL+rpoC1 (74)	0.0	2.53	0.55	0.0	10.13	1.4	0.39-1.73	73.19%

Species discrimination

For the analysis using TaxonDNA the number of correct species identifications using "best match" or "best close match" were favorable in for the majority the five loci or their combinations, as in all the cases the identification success was >60% (Table 5). The

combinations *ycf1*+ITS+*matK* had the highest success rate for the correct identification of species (Best match: 62.16%; Best close match: 62.16%), followed by *matK*+ITS and *ycf1*+ITS+*matK*+*rpoC1*. The combinations *ycf1*+*rpoC1* (Table 5) had the lowest discrimination success rate (Best match: 9.09%; and Best close match: 9.09%). On the basis of "all species barcodes", the identification success was maximum in *ycf1*+*matK* (28.98%) followed by *rpoC1* (23.43%) and *matK* (19.75%) (Table 4). To evaluate efficacy of the genes to produce species-specific clusters, we used the "cluster" function of the TaxonDNA at two different thresholds, 1% and 0.5%. At a 1% threshold, *ycf1* performed better by producing 12 clusters, and 11 of those clusters included only one species (Table 6). At a 0.5% threshold, the combination of *ycf1*+ITS produced the maximum number of clusters (20), but the number of clusters with only one species was 12 (Table 6). For the tree-based analysis, the performance of candidate barcodes at discriminating species were summarized in Figure 1. Tree-based analysis can provide a means for evaluating discriminatory performance by calculating the proportion of monophyletic species in the tree.

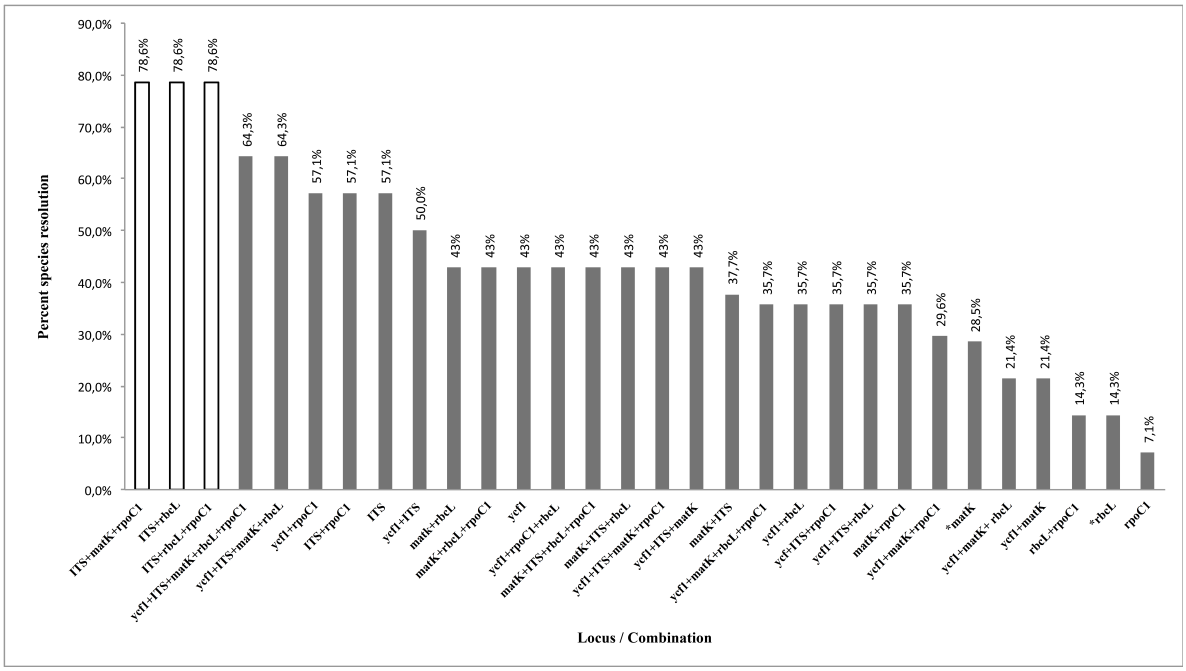
Table 5. Identification success based on the "best match", "best close-match" and "all species barcodes" function of the program TAXON- DNA for *Eucalyptus* species

	Best match			Best close match			All species barcodes		
Barcode locus	**Correct %**	**Ambiguous %**	**Incorrect %**	**Correc %**	**Ambiguous %**	**Incorrect %**	**Correct %**	**Ambiguous %**	**Incorrect %**
ITS	55.00	18.75	26.25	55.00	18.75	26.25	10.00	88.75	1.2
rpoC1	10.93	82.81	6.25	10.93	82.81	6.25	23.43	76.56	0.0
matK	23.45	69.13	7.4	22.22	69.13	6.17	19.75	77.77	0.0
rbcL	16.41	67.16	16.41	14.92	65.67	16.41	13.43	83.58	0.0
ycf1	26.92	7.69	65.38	26.92	7.69	65.38	11.53	88.46	0.0
ITS+rbcL	44.44	8.33	47.22	44.44	8.33	47.22	6.94	93.05	0.0
ITS+rpoC1	33.33	15.27	51.38	33.33	15.27	51.38	1.38	98.61	0.0
rbcL+rpoC1	15.71	78.57	5.71	15.71	78.57	5.71	4.28	95.71	0.0
matK+ITS	61.64	5.47	32.87	61.64	5.47	32.87	15.06	83.56	1.3
matK+rpoC1	13.88	81.94	4.16	13.88	81.94	4.16	13.88	86.11	0.0
matk+rbcL	18.57	50.0	31.42	18.57	50.0	31.42	12.85	87.14	0.0
ycf1+ITS	57.53	8.21	34.24	57.53	8.21	34.24	8.21	91.78	0.0
ycf1+matK	27.53	66.66	5.79	27.53	65.21	5.79	28.98	68.11	1.4
ycf1+rbcL	18.18	63.63	18.18	18.18	63.63	18.18	3.03	96.96	0.0
ycf1+rpoC1	9.09	78.78	12.12	9.09	78.78	12.12	18.18	81.81	0.0
ITS+rbcL+rpoC1	46.26	7.46	46.26	46.26	7.46	46.26	5.97	94.02	0.0
matK+rbcL+rpoC1	16.66	48.61	34.72	16.66	48.61	34.72	6.94	93.05	0.0
ITS+matK+rpoC1	60.27	4.1	35.61	60.27	4.1	35.61	10.95	89.04	0.0
matK+ITS+rbcL	58.9	0.0	41.09	58.9	0.0	41.09	9.58	90.41	0.0
ycf1+ITS+rbcL	52.17	5.79	42.02	52.17	5.79	42.02	10.14	89.85	0.0
ycf+ITS+rpoC1	44.11	16.17	39.7	44.11	16.17	39.7	5.88	94.11	0.0
ycf1+ITS+matK	62.16	5.4	32.43	62.16	5.4	32.43	6.75	91.89	1.3
ycf1+matK+rbcL	16.9	49.29	33.8	15.49	49.29	33.8	8.45	90.14	0.0
ycf1+matK+rpoC1	15.06	80.82	4.1	15.06	80.82	4.1	10.95	89.04	0.0
ycf1+rpoC1+rbcL	16.9	76.05	7.04	15.49	76.05	7.04	4.22	94.36	0.0
matK+ITS+rbcL+rpoC1	57.53	0.0	42.46	57.53	0.0	42.46	5.47	94.52	0.0
ycf1+ITS+matK+rbcL	57.97	0.0	42.02	56.52	0.0	42.02	14.49	84.05	0.0
ycf1+ITS+matK+rpoC1	60.8	4.05	35.13	60.8	4.05	35.13	5.4	94.59	0.0
ycf1+matK+rbcL+rpoC1	16.43	47.94	35.61	15.06	47.94	35.61	5.47	93.15	0.0
ycf1+ITS+matK+rbcL+rpoC1	56.75	0.0	43.24	55.4	0.0	43.24	5.4	93.24	0.0

Table 6. Cluster analysis of *Eucalyptus* species based on the three plant barcodes

Barcode locus	At 1% threshold				At 0.5% threshold			
	No. of clusters	% of clusters with threshold violation	Largest pairwise distance (%)	Clusters with only one species	No. of clusters	% of clusters with threshold violation	Largest pairwise distance (%)	Clusters with only one species
ITS	5	40.0	4.6	3	19	42.1	2.13	11
rpoC1	1	100.0	1.86	0	4	25.0	1.24	3
matK	4	25.0	3.9	3	7	14.28	1.43	5
rbcL	4	25.0	2.05	3	6	16.66	1.92	5
ycf1	12	8.33	2.73	11	16	12.5	1.77	14
ITS+rbcL	2	50.0	10.53	1	5	20.0	10.53	4
ITS+rpoC1	1	100.0	10.5	0	3	33.33	10.5	2
rbcL+rpoC1	1	100.0	3.42	0	1	100.0	3.42	0
matK+ITS	2	50.0	4.7	1	8	37.5	3.14	6
matK+rpoC1	1	100.0	4.41	0	1	100.0	4.41	0
matk++rbcL	1	100.0	4.51	0	4	25.0	4.51	3
ycf1+ITS	5	60.0	6.5	3	20	40.0	2.51	12
ycf1+matK	5	40.0	4.23	4	8	37.5	3.34	6
ycf1+rbcL	5	20.0	6.39	4	7	14.28	6.39	6
ycf1+rpoC1	3	33.33	11.12	2	6	16.66	11.12	5
ITS+rbcL+rpoC1	1	100.0	10.13	0	5	20.0	10.13	4
matK+rbcL+rpoC1	1	100.0	4.51	0	1	100.0	4.51	0
ITS+matK+rpoC1	1	100.0	10.13	0	3	33.33	10.13	2
matK+ITS+rbcL	1	100.0	10.53	0	7	28.57	3.13	6
ycf1+ITS+rbcL	3	33.33	10.53	2	8	12.5	10.53	7
ycf+ITS+rpoC1	2	50.0	11.39	1	8	12.5	11.39	7
ycf1+ITS+matK	3	66.66	4.7	2	9	33.33	3.32	7
ycf1+matK+rbcL	2	50.0	4.95	1	4	40.0	4.51	4
ycf1+matK+rpoC1	2	50.0	5.86	1	2	50.0	5.86	1
ycf1+rpoC1+rbcL	2	50.0	5.97	1	2	50.0	5.97	1
matK+ITS+rbcL+rpoC1	1	100.0	10.13	0	5	20.0	10.13	4
ycf1+ITS+matK+rbcL	3	33.33	10.53	2	8	25.0	2.98	7
ycf1+ITS+matK+rpoC1	2	50.0	10.13	1	4	25.0	10.13	3
ycf1+matK+rbcL+rpoC1	2	50.0	4.51	1	2	50.0	4.51	1
ycf1+ITS+matK+rbcL+rpoC1	2	50.0	10.13	1	6	16.66	10.13	5

According to the results of the NJ cluster analysis, the degree of species resolution was significantly enhanced with the combinations multi-locus regions. Two-region combinations (ITS+*rbcL*) and three-region combinations (ITS+*matK*+*rpoC1* and ITS+*rbcL*+*rpoC1*) resolved 78.6% of the species included in this study (Fig. 1) (Fig. 3a - c). The degree of species resolution for individual barcode regions ranged from 7.1% (*rpoC1*) to 57.1% (ITS*)* (Fig 1.) (Fig. 2a - e).



*Primer developed in this study

Fig. 1 Ability to discriminate (NJ) of 14 species using single and multi-locus combinations. Percent species resolution based on single as well as multi-locus combinations for 14 species of *Eucalyptus*.

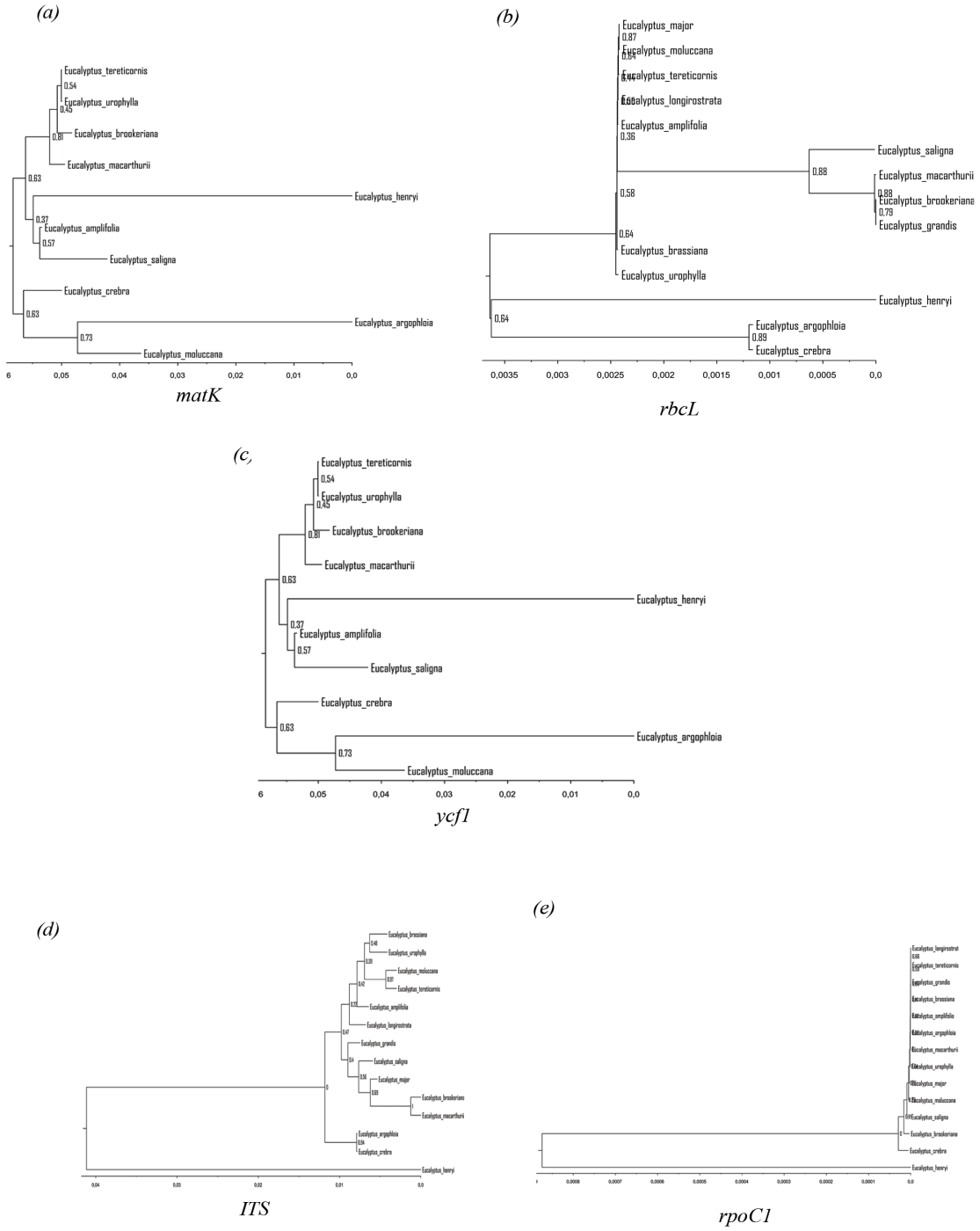


Fig. 2. Single locus Neighbour-joining cluster analysis of 14 *Eucalyptus* species. a, *matK; b, rbcL; c, ycf1; d,* ITS; e *rpoC1***.** The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The trees are drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA 6. Analysis was based on the consensus barcodes.

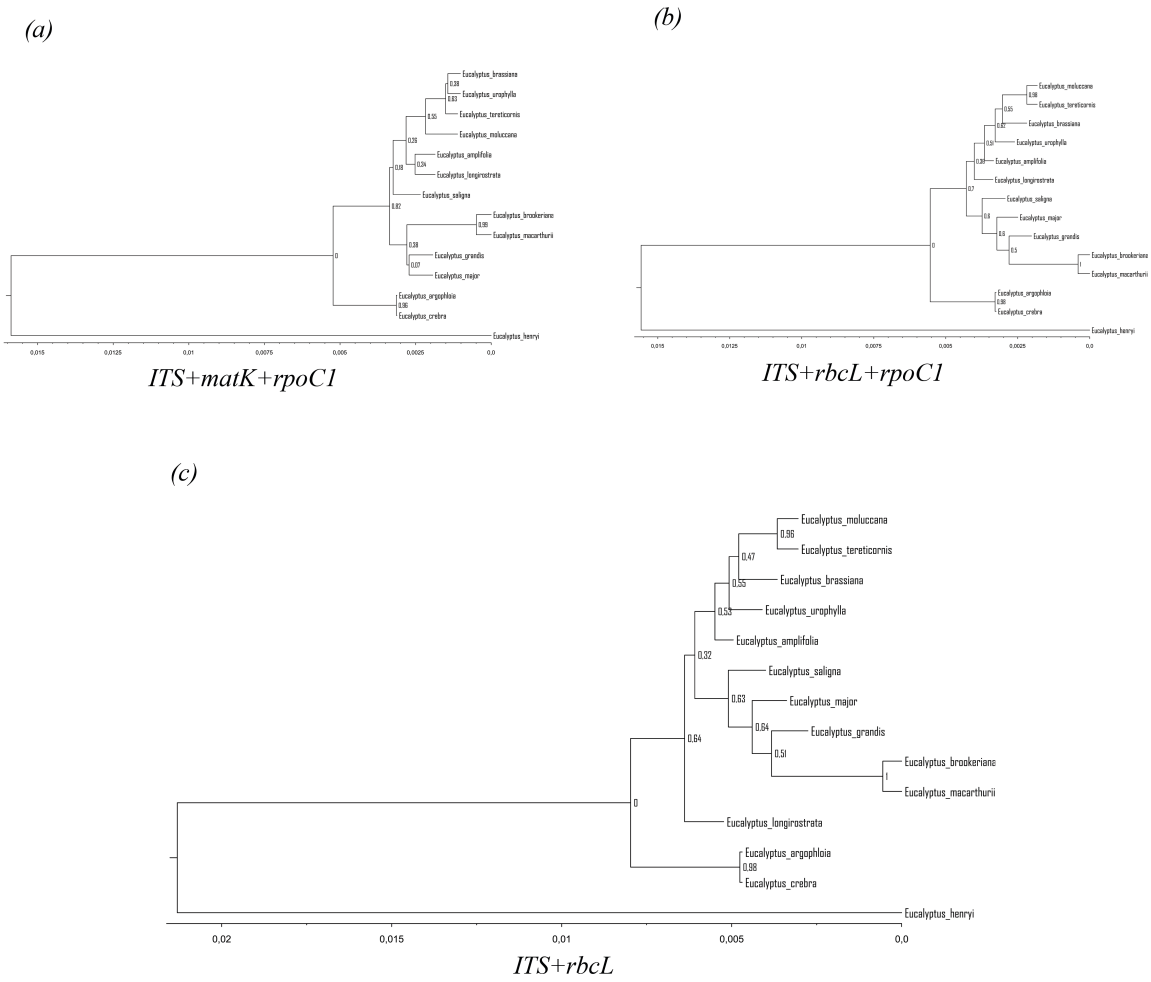


Fig. 3 Neighbour-joining with combined loci cluster analysis of 14 *Eucalyptus* **species.** a, ITS+*matK*+*rpoC1; b,* ITS+*rbcL*+*rpoC1; c,* ITS+*rbcL***.** The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The trees are drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA 6. Analysis was based on the consensus barcodes.

Discussion

Evaluation of the DNA barcodes in *Eucalyptus*

We tested five plant loci as DNA barcodes for species differentiation in the *Eucalyptus* species. Four barcode loci, ITS*, matK*, *rpoC1* and *rbcL* have already been approved and established for plant barcoding (9) and the fifth locus, *ycf1* is being debated for its potential use as a DNA barcode in different plant groups (11). Efficiency of the barcodes in the current studies was judged based on their potential to differentiate among Angiosperms. In our studies, the amplification success of all the three loci was more than 90% for ITS*, matK*, *rpoC1* and *rbcL,* which is comparable with those reported for some other taxa (16,17,26–28), while *ycf1* fragments were only successfully amplified in 67.14% (Table S3), presenting a failure in PCR in some species of *Eucalyptus*, data are similar with Neubig & Abbott (29) which reported a low success of PCR in families *Lauraceae* and *Annonaceae*.

According to our results, ITS and *ycf1* are more parsimony informative sites and better discriminatory power among the five proposed loci, i.e., ITS*, matK*, *rpoC1*, *rbcL* which is consistent with the results of many previous studies (10,29,30). We analyzed the intraspecific and interspecific pairwise distances to determine whether the maximum intraspecific divergence was smaller than the minimum interspecific divergence. The analysis showed an overlap between the intraspecific and interspecific distances in the individual or the combined barcode sequences. Barcode datasets can be used to delimit species by determining the barcode in the distribution of pairwise differences. The distance analysis demonstrated that *ycf1*+ITS+*matK*+*rpoC1* had the highest intraspecific sequence divergence, whereas *ycf1* had the highest interspecific (Table 3). The combination of regions *matK, rbcL* and *ycf1* has been proposed as a universal barcode high resolution at species level in plants (11,31). The discriminatory power of ITS and *matK* lies in the high rate of substitution in the gene,

percentage of variation and the presence of mutationally conserved sectors (32,33). Unfortunately, *matK* can be difficult to PCR amplify using existing primer sets particularly (26), as it happened in this study with primer set reported by Sharma et al. (18). In contrast, the barcode region of *rbcL* and *rpoC1* are easy to amplify, sequence, and align in most land plants and provides utility to the barcode dataset, despite it having only modest discriminatory power (26,34). Therefore, it seems that the rate of correct identification of species using multilocus combinations can increase, in our studies, when we performed intraspecific vs. interspecific sequence divergence analysis, a significant number of sequences overlapped and did not warrant further analysis to determine barcode. Although researchers have used barcode and distance method to distinguish plant groups above the species or generic levels (35), the barcode in ITS*, matK, rbcL*, *rpoC1* and *ycf1* at the species level has been reasonably documented in several plant genera (10,11,36,37).

Although success for DNA barcode plants to be close to 70% (7), new regions are being proposed for greater discrimination of the land species plants (11). In *Eucalyptus* the interspecific pairwise distances in *rpoC1, matK* and *rbcL* were lower as compared to ITS and *ycf1*. This is not surprising as the ITS and *ycf1* region are variable even among organisms of the same species; and a similar pattern of pairwise distance divergence has been found by some other researchers in other plant genera (38–40). We used K2P model to compute pairwise sequence distances. The K2P model considers that transition and transversion occurs at different rates and takes into account both transition and transversion rates to calculate the divergence between sequences (41). This model has been widely used in various investigations for sequence comparison barcode and species identification (10,37). *Eucalyptus* species assignments were made by “best match”, “best close match” and “all species barcodes” functions of TaxonDNA (24). These statistics have been used by a large number of researchers in barcode studies on plants (42,43) for species assignments. In our study the

locus combination *ycf1*+ITS+*matK,* *matK*+ITS*,* *ycf1*+ITS+*matK*+*rpoC1* and ITS+*matK*+*rpoC1* showed a higher intraspecific variation, being consistent with data from "best match" and “best close match” above 60%. According to Ashfaq et al. (44) resolution identification of species of cotton was higher when using multilocus combination. Xu et al. (10) confirmed DNA barcoding to differentiate *Dendrobium* species and found *matK*+ITS the most effective barcode to identify these species. Several studies have demonstrated the power of ITS for discrimination between species of terrestrial plants (26,37,45).

We performed cluster analysis to evaluate the efficiency of barcodes to separate the species, functions of TaxonDNA. At a threshold level of 0.5% *ycf1* was the most successful barcode, producing 16 cluster, and 14 clusters with only one species, followed by ITS producing 19 clusters and 11 clusters with only one species. The locus combinations reduce the formation of groups with single-species profiles. For example (*matK*+ITS+*rbcL*+*rpoC1*) producing five clusters, and four clusters with a single species. This analysis helps us understand the power of resolution taking into account the feasibility of the threshold values to distinguish intra- and interspecific variability by 0.5%, and 1% thresholds. When species are grouped only by a cluster, it shows little variation of the barcode region, however as it increases the number of cluster, the resolution of species is higher. In our study, the tree-based method outperformed the similarity-based methods. We evaluated the species resolution abilities of the five DNA barcode regions and their combinations using a tree-based (NJ) method and revealed ITS+*rbcL*, ITS+*matK*+*rpoC1* and ITS+*rbcL*+*rpoC1* are the best combinations DNA barcode affording 78.6% species resolution, thus apparently pointing towards its suitability as one of the candidate DNA barcodes for land plants. The degree of species resolution for individual barcode regions ranged from 7.1% to 57.1% (*rpoC1* and ITS, respectively; Fig. 1). The NJ tree based on the locus ITS returned the best resolution based only in one region and previously, it has been reported that this region presented a higher discrimination power than

the other datasets to identify *Passiflora* species (43). Chen et al. (17) in a study on identifying medicinal plant species determined that the *ITS2* region provided the highest species resolution in a 97.7%. In another study Hartvig et al. (37), have reported that nuclear marker (ITS), chloroplast markers (*matK* and *rbcL*) and combinations showed a high rate of greater discrimination 90% for species of *Dalbergia*. The NJ tree based on the *ycf1* is shown in Fig. 2 - c, the degree of species resolution was from 43%, this marker has been proposed as the most promising single-locus barcode for land plants (11). However, further studies are required in the design of a new primers for *ycf1* region from sequences of *Eucalyptus*, this will allow greater in the percentage PCR success, which may allow better discrimination of *Eucalyptus* species. For the others regions, this success were lower than 30% and only combined with other regions it can reach a reasonable resolution. Several combinations of two or three barcodes have been proposed as core barcodes, including ITS2+*matK* (44), *rbcL+trnH-psbA+matK* (46), *ycf1* (29), *trnH-psbA+rbcL+matK* (30), *rbcL+matK+trnH-psbA+nrITS* (47), *rbcL+matK* (48), *ycf1b+matK+rbcLb* (11), however, a consensus has not been reached (49). The success of ITS+*rbcL*, ITS+*matK+rpoC1* and ITS+*rbcL+rpoC1* as a core barcode was based on the straightforward recovery of the *rbcL* region and the discriminatory power of the *matK* and ITS region. The region barcode *matK* is one of the most rapidly evolving coding sections of the plastid genome, and is perhaps the closest plant analogue to the *CO1* animal barcode, this barcode offers higher level of variation between species in animal. However in land plants, the plastid DNA regions *matK* and *rbcL* do not offer such a barcode, particularly in closely related species, being necessary the use of other more variable regions to assess greater resolutions (50). A possible limitation in using these three regions is the difficulty in alignment due to the considerable length variations caused by indels and simple sequence repeats indeed, a few manual alignment adjustments were necessary in this study. Find a unique barcode for *Eucalyptus* has proved a difficult task. It is required the combination of

nuclear and plastid genomes for discrimination of most species of *Eucalyptus*. However, considering the cost-benefit, we suggests the use of two-locus region (ITS+*rbcL*) for the species identification in this genus.

Conclusions

Our study shows that a combination of ITS+*rbcL* is most suitable for barcoding the *Eucalyptus* species separating *E. moluccana, E. tereticornis, E. urophylla, E. amplifolia, E. saligna, E. major, E. grandis, E. brookeriana, E. macarthurii, E. longirostrata* and *E. Corymbia henryi,* by tree-based (NJ) method. These combinations provide the highest degree of species resolution, considering the other combinations.

References

1. Soares NS, Silva ML da, Rezende JLP de, Gomes MFM. Competitividade da cadeia produtiva da madeira de eucalipto no Brasil. Rev Árvore. 2010;34(5):917–28.

2. Filho EP, Santos TDPE. Program de melhoramento genético de eucalipto da Embrapa Florestas: resultados e perpectivas. Embrapa Florestas. 2011;214.

3. Pongitory V, Moura G, Blume E, Flores E, Jaya DI. O germoplasma de *Eucalyptus urophylla* S. T. Blake no Brasil. Comun técnico. 2004; (Blake 1977).

4. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc Biol Sci. 2003;270(1512):313–21.

5. Ratnasingham S, Hebert PDN. The Barcode of Life Data System BOLD. Mol Ecol Notes [Internet]. 2007;7(3):355–64. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?doi=10.1111/j.1471-8286.2007.01678.x

6. Ferri G, Corradini B, Ferrari F, Santunione a. L, Palazzoli F, Alu' M. Forensic botany II, DNA barcode for land plants: Which markers after the international agreement? Forensic Sci Int Genet [Internet]. Elsevier Ireland Ltd; 2015;15:131–6. Available from: http://linkinghub.elsevier.com/retrieve/pii/S187249731400221X

7. CBOL. A DNA barcode for land plants. Proc Natl Acad Sci U S A [Internet]. 2009;106(31):12794–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3241790&tool=pmcentrez&rendertype=abstract

8. Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, et al. Selecting barcoding loci for plants: Evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol Ecol Resour. 2009;9:439–57.

9. Hollingsworth PM. A DNA barcode for land plants. Mol Ecol Resour [Internet]. 2014 May;14(3):437–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24286499

10. Xu S, Li D, Li J, Xiang X, Jin W, Huang W, et al. Evaluation of the DNA barcodes in dendrobium (Orchidaceae) from mainland Asia. PLoS One. 2015;10(1):1–12.

11. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, et al. ycf1, the most promising plastid DNA barcode of land plants. Sci Rep [Internet]. 2015;5:8348. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25672218\n. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4325322/pdf/srep08348.pdf

12. Lucas EJ, Belsham SR, Lughadha EMN, Orlovich D a., Sakuragui CM, Chase MW, et al. Phylogenetic patterns in the fleshy-fruited Myrtaceae ? preliminary molecular evidence. Plant Syst Evol [Internet]. 2005 Feb 16 [cited 2014 Jul 24];251(1):35–51. Available from: http://link.springer.com/10.1007/s00606-004-0164-9

13. Steane DA, Nicolle D, Potts BM. Phylogenetic positioning of anomalous eucalypts by using ITS sequence data. Aust Syst Bot. 2007;20(5):402–8.

14. Ochieng JW, Henry RJ, Baverstock PR, Steane D a, Shepherd M. Nuclear ribosomal pseudogenes resolve a corroborated monophyly of the eucalypt genus Corymbia despite misleading hypotheses at functional ITS paralogs. Mol Phylogenet Evol [Internet]. 2007 Aug [cited 2014 Jul 31];44(2):752–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17570687

15. Doyle J, Doyle J. Isolation of plant DNA from fresh tissue. Focus (Madison). 1990;12:13–5.

16. Tokuoka T, Tobe H. Phylogenetic analyses of Malpighiales using plastid and nuclear DNA sequences, with particular reference to the embryology of Euphorbiaceae sens. str. J Plant Res [Internet]. 2006 Nov [cited 2014 Aug 5];119(6):599–616. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16937025

17. Chen S, Yao H, Han J, Liu C, Song J, Shi L, et al. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS One [Internet]. 2010 Jan [cited 2014 Jul 20];5(1):e8613. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2799520&tool=pmcentrez&rendertype=abstract

18. Sharma A, Folch JL, Cardoso-Taketa A, Lorence A, Villarreal ML. DNA barcoding of the Mexican sedative and anxiolytic plant Galphimia glauca. J Ethnopharmacol [Internet]. 2012 Nov 21 [cited 2014 Aug 5];144(2):371–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23010364

19. Hall T. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser [Internet]. 1999;41:95–8.

Available from: http://jwbrown.mbio.ncsu.edu/JWB/papers/1999Hall1.pdf

20. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res [Internet]. 2004;32(5):1792–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15034147

21. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3-new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):1–12.

22. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30(12):2725–9.

23. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. J Mol Evol. 1980;16(2):111–20.

24. Meier R, Shiyang K, Vaidya G, Ng PKL. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst Biol. 2006;55(5):715–28.

25. Felsenstein J. Phylogenies and the Comparative Method. The American Naturalist. 1985. p. 1–15.

26. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. PLoS One [Internet]. 2011 Jan [cited 2014 Jul 11];6(5):e19254. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3102656&tool=pmcentrez&rendertype=abstract

27. Forrest A, Hollingsworth PP. Plant DNA Barcoding using matK some work on new primer sets. 2011;

28. Kress WJ, Garcı C, Uriarte M, Erickson DL. DNA barcodes for ecology , evolution , and conservation. 2015;30(1).

29. Neubig KM, Abbott JR. Primer development for the plastid region YCF1 in annonaceae and other magnoliids. Am J Bot. 2010;97(6):52–5.

30. Li D-Z, Gao L-M, Li H-T, Wang H, Ge X-J, Liu J-Q, et al. From the Cover: Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc Natl Acad Sci. 2011;108(49):19641–6.

31. Dong W, Liu H, Xu C, Zuo Y, Chen Z, Zhou S. A chloroplast genomic strategy for designing taxon specific DNA mini-barcodes: a case study on ginsengs. BMC Genet [Internet]. 2014;15(1):138. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25526752

32. Hilu K, Liang H. The matK gene: sequence variation and application in plant systematics. Am J Bot. 1997;84(6):830.

33. Selvaraj D, Shanmughanandhan D, Sarma RK, Joseph JC, Srinivasan R V, Ramalingam S. DNA barcode ITS effectively distinguishes the medicinal plant Boerhavia diffusa from its adulterants. Genomics Proteomics Bioinformatics [Internet]. Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China; 2012 Dec [cited 2014 Jul 12];10(6):364–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23317705

34. Roy S, Tyagi A, Shukla V, Kumar A, Singh UM, Chaudhary LB, et al. Universal plant DNA barcode loci may not work in complex groups: a case study with Indian berberis species. PLoS One [Internet]. 2010 Jan [cited 2014 Jul 17];5(10):e13674. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2965122&tool=pmcentrez&rendertype=abstract

35. Zhang CY, Wang FY, Yan HF, Hao G, Hu CM, Ge XJ. Testing DNA barcoding in closely related groups of Lysimachia L. (Myrsinaceae). Mol Ecol Resour. 2012;12(1):98–108.

36. Singh H, Parveen I, Raghuvanshi S, Babbar SB. The loci recommended as universal barcodes for plants on the basis of floristic studies may not work with congeneric species as exemplified by DNA barcoding of Dendrobium species. BMC Res Notes [Internet]. 2012;5(1):42. Available from: http://www.biomedcentral.com/1756-0500/5/42

37. Hartvig I, Czako M, Kjaer ED, Nielsen LR, Theilade I. The use of DNA barcoding in identification and conservation of rosewood (Dalbergia spp.). PLoS One. 2015;10(9).

38. Yang JB, Wang YP, Möller M, Gao LM, Wu D. Applying plant DNA barcodes to identify species of Parnassia (Parnassiaceae). Mol Ecol Resour. 2012;12(2):267–75.

39. Stech M, Veldman S, Larraín J, Muñoz J, Quandt D, Hassel K, et al. Molecular species delimitation in the Racomitrium canescens complex (Grimmiaceae) and implications for DNA barcoding of species complexes in mosses. PLoS One [Internet]. 2013 Jan [cited 2014 Aug 5];8(1):e53134. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3544804&tool=pmcentrez&rendertype=abstract

40. Zeng L, Zhang Q, Sun R, Kong H, Zhang N, Ma H. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat Commun [Internet]. Nature Publishing Group; 2014;5:4956. Available from: http://www.nature.com/doifinder/10.1038/ncomms5956

41. Collins RA, Boykin LM, Cruickshank RH, Armstrong KF. Barcoding's next top model: An evaluation of nucleotide substitution models for specimen identification.

Methods Ecol Evol. 2012;3(3):457–65.

42. Han J, Zhu Y, Chen X, Liao B, Yao H, Song J, et al. The short ITS2 sequence serves as an efficient taxonomic sequence tag in comparison with the full-length ITS. Biomed Res Int. 2013;2013:3–10.

43. Giudicelli GC, Mader G, de Freitas LB. Efficiency of ITS sequences for DNA barcoding in Passiflora (Passifloraceae). Int J Mol Sci. 2015;16(4):7289–303.

44. Ashfaq M, Asif M, Anjum ZI, Zafar Y. Evaluating the capacity of plant DNA barcodes to discriminate species of cotton (Gossypium: Malvaceae). Mol Ecol Resour. 2013;13(4):573–82.

45. Bhagwat RM, Dholakia BB, Kadoo NY, Balasundaran M, Gupta VS. Two new potential barcodes to discriminate Dalbergia species. PLoS One. 2015;10(11):1–18.

46. Kress WJ, Erickson DL, Jones FA, Swenson NG, Perez R, Sanjur O, et al. Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proc Natl Acad Sci U S A. 2009;106:18621–6.

47. Liu J, Möller M, Gao LM, Zhang DQ, Li DZ. DNA barcoding for the discrimination of Eurasian yews (Taxus L., Taxaceae) and the discovery of cryptic species. Mol Ecol Resour. 2011;11:89–100.

48. Saarela JM, Sokoloff PC, Gillespie LJ, Consaul LL, Bull RD. DNA barcoding the Canadian Arctic flora: core plastid barcodes (rbcL + matK) for 490 vascular plant species. PLoS One [Internet]. 2013;8(10):e77982. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3865322&tool=pmcentrez&rendertype=abstract

49. Yao H, Song J, Liu C, Luo K, Han J, Li Y, et al. Use of ITS2 Region as the Universal

DNA Barcode for Plants and Animals. PLoS One [Internet]. 2010;5(10):e13102. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2948509&tool=pmcentrez&rendertype=abstract

50. Jiang Y, Ding C, Zhang L, Yang R, Zhou Y, Tang L. Identification of the genus Epimedium with DNA barcodes. J Med Plants Res. 2011;5(28):6413–7.

Supporting Information

Table S1. The 34 plastid genomes used for *matK* and *rbcL* primer design

Species	**Accession Number**	**Version**
Eucalyptus grandis	NC_014570.1	GI:309322431
Eucalyptus curtisii	NC_022391.1	GI:545717342
Eucalyptus erythrocorys	KC180799.1	GI:442568712
Eucalyptus cladocalyx	NC_022394.1	GI:545717600
Eucalyptus melliodora	KC180784.1	GI:442567422
Eucalyptus melliodora	NC_022392.1	GI:545717428
Eucalyptus polybractea	NC_022393.1	GI:545717514
Eucalyptus deglupta	NC_022399.1	GI:545718030
Eucalyptus grandis	HM347959.1	GI:308223265
Eucalyptus camaldulensis	NC_022398.1	GI:545717944
Eucalyptus saligna	NC_022397.1	GI:545717858
Eucalyptus aromaphloia	NC_022396.1	GI:545717772
Eucalyptus globulus	KC180787.1	GI:442567680
Eucalyptus globulus subsp. globulus	NC_008115.1	GI:108802622
Eucalyptus salmonophloia	NC_022403.1	GI:545718374
Eucalyptus diversicolor	NC_022402.1	GI:545718288
Eucalyptus torquata	NC_022401.1	GI:545718202
Eucalyptus spathulata	NC_022400.1	GI:545718116
Eucalyptus nitens	NC_022395.1	GI:545717686
Eucalyptus marginata	NC_022390.1	GI:545717256
Eucalyptus microcorys	NC_022404.1	GI:545718460
Eucalyptus guilfoylei	KC180798.1	GI:442568626
Eucalyptus regnans	NC_022386.1	GI:545716912
Eucalyptus elata	NC_022385.1	GI:545716826
Eucalyptus sieberi	NC_022384.1	GI:545716740
Eucalyptus verrucata	NC_022381.1	GI:545716482
Eucalyptus patens	NC_022389.1	GI:545717170
Eucalyptus radiata	NC_022379.1	GI:545716310
Eucalyptus delegatensis	NC_022380.1	GI:545716396
Eucalyptus obliqua	NC_022378.1	GI:545716224
Eucalyptus cloeziana	NC_022388.1	GI:545717084
Eucalyptus umbra	NC_022387.1	GI:545716998
Eucalyptus baxteri	NC_022382.1	GI:545716568
Eucalyptus diversifolia	NC_022383.1	GI:545716654

Table S2. Evaluation of five DNA markers and combinations of the markers

Barcode locus	Universality of primers	Percentage PCR success (%)	Percentage sequencing success (%)	Length of aligned sequence (bp)	No. of parsimony informative sites/variable sites	No. of species samples (individuals)	Ability to discriminate (NJ)
ITS	Yes	100	87.14	800	94/138	81	64%
rpoC1	Yes	100	88.57	594	8/29.0	64	28.57%
**matK*	Yes	100	87.14	892	32/85	81	85%
**rbcL*	Yes	100	80	840	25/77	67	85%
ycf1	Partial	67.14	40	663	71/190	26	80%
ITS+rbcL	-	-	-	1640	120/211	72	92%
ITS+rpoC1	-	-	-	2286	134/244	72	71.42%
rbcL+rpoC1	-	-	-	1434	32/105	70	92.85%
matK+ITS	-	-	-	1692	127/216	73	57%
matK+rpoC1	-	-	-	1486	39/110	72	64%
matk+rbcL	-	-	-	1732	57/159	70	100%
ycf1+ITS	-	-	-	1463	166/324	73	50%
ycf1+matK	-	-	-	1555	103/272	69	42.85%
ycf1+rbcL	-	.	-	1503	96/267	66	50%
ycf1+rpoC1	-	-	-	1257	78/218	66	78.57%
ITS+rbc+rpoC1	-	-	-	2234	126/235	67	92%
matK+rbcL+rpoC1	-	-	-	2336	64/187	72	100%
ITS+matK+rpoC1	-	-	-	2286	134/244	73	42%
matK+ITS+rbc	-	-	-	2532	152/293	73	50%
ycf1+ITS+rbcL	-	-	-	2303	191/397	69	35.71%
ycf+ITS+rpoC1	-	-	-	2057	171/347	68	42.85%
ycf1+ITS+matK	-	-	-	2355	198/406	74	42%
ycf1+matK+ rbcL	-	-	-	2395	128/349	71	50%
ycf1+matK+rpoC1	-	-	-	2149	110/300	73	57.14%
ycf1+rpoC1+rbcL	-	-	-	2097	103/295	71	57.14%
matK+ITS+rbcL+rpoC1	-	-	-	3126	159/321	73	0.0%
ycf1+ITS+matK+rbcL	-	-	-	3195	222/477	69	78%
ycf1+ITS+matK+rpoC1	-	-	-	2949	205/434	74	50%
ycf1+matK+rbcL+rpoC1	-	-	-	2989	135/377	73	64%
ycf1+ITS+matK+rbcL+rpoC1	-	-	-	3789	230/511	74	78.57%